26.8 The Asynchronous 24MB On-Chip Level-3 Cache for a Dual-Core Itanium®-Family Processor

John Wuu, Don Weiss, Charles Morganti, Michael Dreesen

Intel, Fort Collins, CO

With smaller device sizes and larger die areas than before, recent microprocessors have incorporated increasingly larger caches onto their dies [1-2]. This next-generation, dual-core Itanium® processor, built on a 90nm 7M process, incorporates 24MB of on-chip Level 3 (L3) cache [3]. The cache dedicates 12MB to each core, and each single-ported, 12-way set associative cache has a 128B line size. An asynchronous design style provides many advantages in addressing the challenges that arise from designing such a large cache, which totals more than 1.47B transistors.

To efficiently utilize area on the die and to allow for flexible floor planning, this cache continues to employ the subarray-style design introduced in the second generation Itanium® processor [4]. In addition, every 12 subarrays, each belonging to a different way, is arranged into a cluster, and 32 data, 1 redundant, and 2 ECC clusters make up a 12MB cache (Fig. 26.8.1). All clusters are logically identical but may differ physically (Fig. 26.8.7). The subarrays are arranged in a grid-like manner providing repeater channels between every subarray column and every four subarray rows. The repeater channels reduce the overall array efficiency to 78% but their regularity enables automated cache composition, including repeater and route insertion. After a floor plan change, the automated flow completely recomposes the cache in less than two days.

Reducing cache power is important in large, power limited processor designs. Cache active power is reduced by traditional bank organization, activating one of 12 subarrays per access and by the asynchronous design that eliminates all clock distribution and latch power. Using non-minimal length devices in peripheral circuitry to decrease sub-threshold current helps reduce the standby power but gate leakage dominates in SRAM cells. Therefore, the L3 cache is placed on a separate variable supply that uses on-chip circuitry to measure both core and cache power and dynamically varies each voltage [3]. The cache voltage tracks ~300mV below the core voltage to trade off power and performance. As a result, this 24MB cache consumes <4.2W, with 95% due to leakage, while the cores operate nominally at 1.1V.

With larger multi-cycle on-chip caches, traditional synchronous design becomes increasingly inefficient. Much of the total delay is dedicated to clock skew, latch delay, margin in each cycle, and non-ideal division to cycle boundaries. In addition, significant margin must be added to the SRAM cell access cycle to account for slow, marginal cells that are statistically probable in a 24MB cache. The delivery of low skew clocks over such a large area is also difficult and costly. As shown in Fig. 26.8.2, a traditional synchronous design with the lowered supply would result in eight cycles of latency through the data arrays. This single-ended asynchronous design eliminates the drawbacks above and achieves a 5-cycle array access, resulting in a 14-cycle load-use latency.

For a read access, addresses are sent from the datapath to the subarrays and decoded by static first-level decoders in the mid-logic block (Fig. 26.8.3). Addresses are sent in 1-hot complimentary pairs to prevent glitches on decoder outputs due to address path variations. The first-level decoder outputs are sent to, and further decoded in, each of the 16 groups to generate WLs and column select signals. Figure 26.8.4 illustrates the single-ended sensing unit similar to the one used in the Itanium 2® processor [4]. Improvements, such as connecting column prechargers to column selects and replacing the dynamic sensing logic with a static NAND gate, are made to offer better leakage immunity. The analog signal (awwtm) used to control Programmable Weak Write Test Mode (PWWTM) [5] defaults to 0 in normal operation but may be set to non-zero values to weaken the holder and trade off leakage protection for speed.

The subarray outputs, precharged at the beginning of a read access, are wire-OR'ed within a cluster to accomplish bank multiplexing (Fig. 26.8.5). Each cluster outputs 32b from the active subarray, and a total of 1024 data, 40 ECC, and 32 redundancy bits are returned to the synchronous L3 datapath unit. The data flows through a transparent latch before proceeding to redundancy and chunk muxes, after which a 256b data chunk and its 10 ECC bits are returned to the core every cycle for four cycles,starting with the critical chunk. The cache sustains consecutive read or write operations every four cycles, and interleave a pair of read and write operations every five cycles. Figure 26.8.6 depicts the critical path described above, accessing the subarray farthest from the datapath.

To improve yield and reliability, three approaches are taken. To address manufacturing defects, shifted-block redundancy is employed. Defects are replaced in half-subarray increments, with one redundant subarray provided for each 1MB way. Irreparable ways are disabled for separate product bins with reduced cache. The cache is protected from soft errors by ECC capable of correcting one error and detecting two errors for every 256b chunk. Finally, to address latent defects ECC errors are monitored by Pellston technology. If a line reports an ECC error, firmware writes back the corrected data and attempts the read again. A consecutive ECC error forces the line into a "disable" MESI state. The number of lines Pellston disables is programmed in firmware and is set small enough to ensure negligible performance impact.

L3 test features include Programmable Built-In Self Test (PBIST) and Direct Access Test (DAT) to allow access to the cache through the I/O pins. Key signals in subarrays are observed through JTAG. A separate test block is used to calibrate PWWTM which detects retention faults. Finally, two L3 tag ways are configured to capture chip level signals for chip and system level debugging.

The cache is characterized to operate above 2.0GHz at 0.8V. At 2.0GHz, interleaving read and write operations achieves a maximum bandwidth of 102.4GB/s.

Acknowledgements:
The authors thank L. Cluff, S. Liepe, T. Miles, S. Stevens, M. Unangst, and R. Woodruff for their contributions.

References:
[1] D. Wendell et al., "A 4MB On-Chip L2 Cache for a 90nm 1.6GHz 64b SPARC Microprocessor," *ISSCC Dig. Tech. Papers,* pp. 66-67, Feb., 2004.
[2] J. Chang et al., "A 0.13um Triple-Vt 9MB Third Level On-Die Cache for the Itanium 2 Processor," *ISSCC Dig. Tech. Papers,* pp. 496-497, Feb., 2004.
[3] S. Naffziger et al., "The Implementation of a 2-Core, Multi-Threaded Itanium®-Family Processor," *ISSCC Dig. Tech. Papers*, Paper 10.1, pp. 182-183, Feb., 2005.
[4] D. Weiss et al., "The On-Chip 3MB Subarray Based 3rd Level Cache on an Itanium Microprocessor," *ISSCC Dig. Tech. Papers,* pp. 112-113, Feb., 2002.
[5] S. Rusu et al., "A 1.5-GHz 130-nm Itanium 2 Processor With 6-MB On-die L3 Cache," *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1887-1895, Nov., 2003.

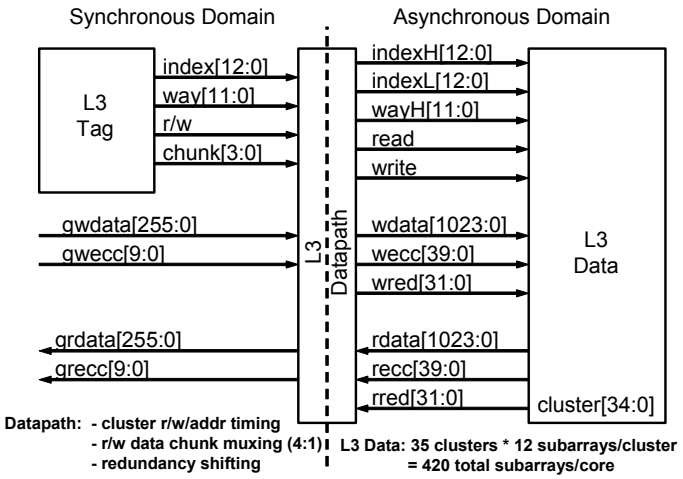

Figure 26.8.1: L3 cache block diagram.

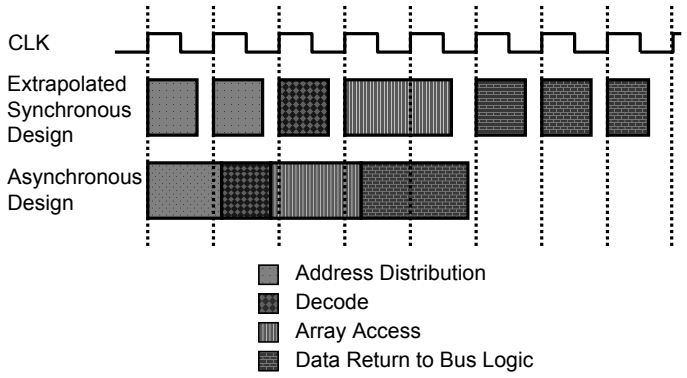

Figure 26.8.2: Timing comparison.

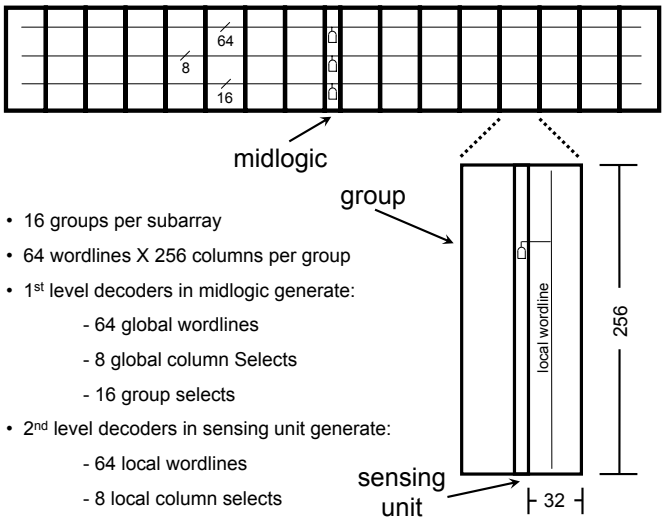

Figure 26.8.3: Subarray organization.

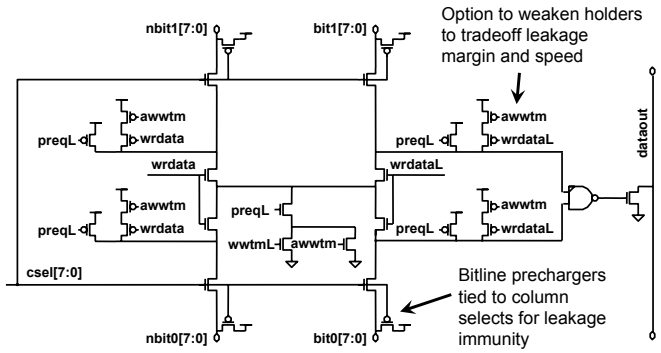

Figure 26.8.4: Sensing unit.

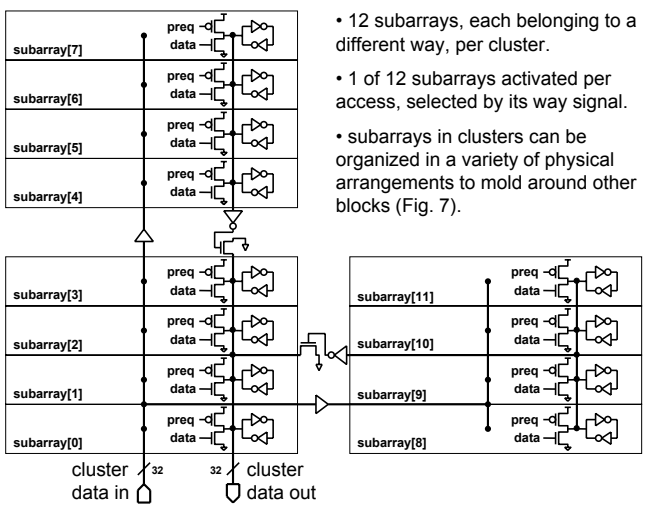

Figure 26.8.5: Bank multiplexing in sample cluster.

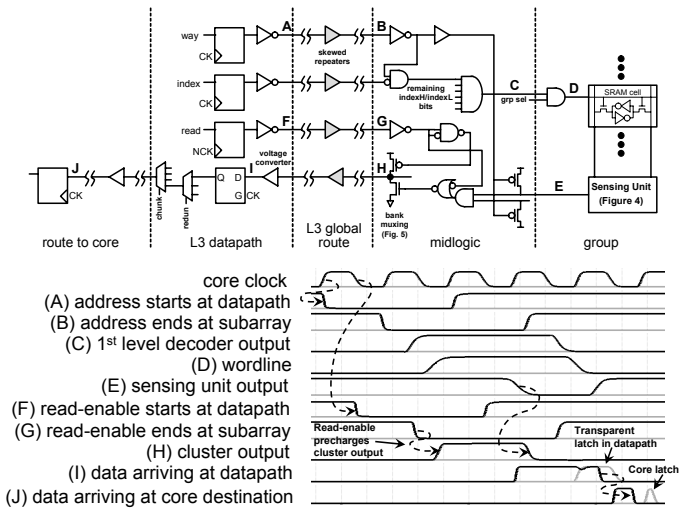

Figure 26.8.6: Read path spice waveforms.

Continued on Page 612

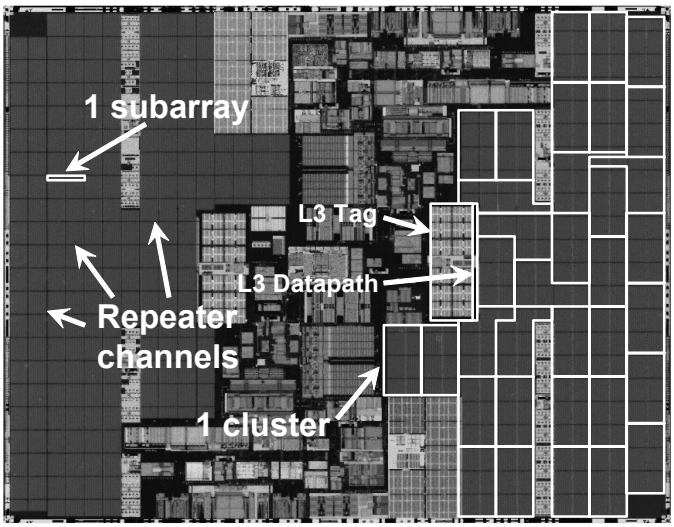

Figure 26.8.7: Die photomicrograph.

